Applied Research on English Language

V. 12 N. 3 2023
pp: 67-86
http://jare.ui.ac.ir

DOI: 10.22108/are.2023.138647.2129
Document Type: Research Article

Leniency in Translation Assessment and Raters' Personality Traits of Agreeableness and Conscientiousness

Zohreh Gharaei *

Assistant Professor, English Department, Faculty of Literature and Foreign Languages, University of Kashan, Kashan, Iran

Received: 2023/08/07 **Accepted: 2023/12/09**

Abstract: Considering the body of research on the intersection of psychology and assessment that suggests assessment is not only under the influence of the model against which a translation is assessed, the present study aimed to investigate the role of raters' personality traits in their assessment. It was also intended to see if the application of an agreed-upon unified assessment tool could change the way evaluations are made, and if this possible change is the same among raters with different personality traits. To this end, seven raters were asked to score an English text translated by 23 students twice with an interval of at least two weeks. In the first scoring, they were asked to assess the way they usually did with no specific guidelines. However, the second assessment was done through the application of an agreed-upon model. Moreover, the raters were scored on the personality traits of agreeableness and conscientiousness as it is believed that these two traits are predictors of leniency. The data analyzed revealed (1) a significant correlation between the raters' score on agreeableness and the scores they assigned to the translations, and (2) a significant negative correlation between the raters' score on conscientiousness and the scores they assigned to the translations. This proved that more agreeable less conscientious raters are prone to be more lenient. The value of Cronbach alpha also revealed that more conscientious less agreeable raters were more consistent in their assessments in both rounds of scoring. This, however, turned out not to be the case for less conscientious raters.

Keywords: Translation Assessment, Agreeableness, Conscientiousness, Leniency, Model.

* **Corresponding Author.**

Authors' Email Address:
[1] **Zohreh Gharaei** (zgharaei@kashanu.ac.ir)

Introduction

Having read a translation, having a text translated, or having the experience of translation even once in a lifetime gives one an impetus to foster the concern of judging the quality of translation. This concern has occupied the minds of many from time immemorial and systematic efforts have been made to make translation quality assessment (TQA) as objective as possible. However, to some translation is an individual and creative act that depends on the interpretations, decisions, and intuitions of the translator and, therefore, a subjective activity in nature. This reflects a hermeneutic or neo-hermeneutic attitude (House, 1997; 2001). While this position might not be challenged per se, it is in essence in clash with the task of assessment.

To base our assessment on solid grounds, we inevitably need to resort to a theory of meaning with more scientific rigor and fewer degrees of relativity. Given the rise of Translation Studies (TS) as an academic discipline, the need is felt more than ever before. For a discipline to be regarded as scientific, training is one of the requirements, and assessment is so closely connected with training that it is not possible to work properly in one area without being constantly involved with the other (Heaton, 1990). In response to this challenge, some have directed their attention to the development of a valid, reliable model of assessment. However, a review of the existing literature reveals that despite all attempts made in recent decades to develop models for TQA, there is still no agreed-upon model that addresses all concerns. Besides, research has revealed that the personality of the rater affects the quality of the assessments. This could be extended to TQA (Bernardin, Tyler, & Villanova, 2009; Buchanan, 2017; Conde, 2012; Li & Yang, 2023; Tziner et al., 2005). Therefore, even if an agreed-upon model existed or could be developed to be used by different raters assessing the very same translated work, they would not necessarily end up with the same judgment and it would not be an ending point to the disparities. While it is not just the model that affects the outcome of translation quality assessment and the role of other factors has been confessed, such factors have received scarce attention in research. Against this backdrop, in the present study, the relationship between raters' personality traits and their TQA is investigated. Moreover, it is to test if the use of an agreed-upon model affects the quality assessment of raters with different personality traits.

Conducting such studies is important in improving educational practice. To be more specific, if personality traits can potentially introduce biases in the evaluation process, the objectivity of the assessments could be called into question. Therefore, to enhance the validity of the assessment, the effect of such factors should be identified. Moreover, understanding the influence of personality traits on quality assessment can improve rater training.

Background

In what follows, a review of the existing academic theories and models of TQA is presented along with the criticisms leveled at each. In light of the discussion put forward, it is aimed to highlight the possible shift of direction whose traces are observable in more recent research in TQA.

The Models: Pros and Cons

In the pre-linguistic era of TS, evaluative language was vague and relied heavily on concepts such as "smooth," "natural," "good," and "faithful" on one end, and "awkward," "artificial," "bad," and "betraying" on the other end (Munday, 2016). Mindful of this, Nida (1964) attempted to introduce a scientific framework for the translation process and more objective evaluation criteria by proposing the principle of equivalent effect. However, this principle faced significant criticism due to its subjective nature (House, 2001; Munday, 2016). Other scholars, such as Carrol and Reiss, have presented alternative approaches to TQA. Carrol (1966) focused on measuring the quality of technical and scientific texts based on intelligibility and informativeness, while Reiss (1989) formulated a model that emphasized the influence of text type on translation strategies and evaluation criteria.

Advocating a more systematic, linguistically-based approach, House (1997; 2009) distinguished between mentalist, response-based, and discourse approaches. Her model, rooted in Hallidayan systemic-functional theory, addressed elements of register and employed comparative analysis of source text (ST) and target text (TT) to identify errors and categorize them. However, critics have highlighted that her model is demanding, time-consuming, and complex, restraining its practical application in educational or workplace settings (Drugan, 2013; Williams, 2001). Among other scholars in the field, one can refer to Williams (2001; 2005; 2009) and his attempt to propose a philosophical, discourse-based model drawing on Toulmin's (1964; 1984) argumentation theory. This model received positive feedback for its consideration of real-world factors, utilization of longer and more diverse sample texts, and the inclusion of an evaluation grid. Nonetheless, critics have pointed out that the model is time-consuming and fails to consider the translation process (Drugan, 2013; Williams, 2001).

There have also been proposals for eclectic evaluation tools in the literature. A case in point is Al-Qinai's (2000) empirical model which involves comparing translations with the original text across seven parameters of textual typology. He applied the model to analyze two short advertisements and the outcome was a qualitative description of each parameter.

However, Drugan (2013) raises valid concerns about the small sample texts. The time-consuming nature of the model also makes it less practical. Another significant criticism aimed at Al-Qinai's model is the lack of pass or fail criteria, as well as the absence of scores or weights.

The literature reveals that even the most popular existing models have failed to meet the demands of real-world situations and lack practicality. Drugan (2013) emphasizes the divide between theorists and professionals, as evidenced in interviews and research visits to Language Service Providers (LSPs). Interestingly, not a single academic model was mentioned as a means of assessing translation quality in real-world contexts. These findings highlight the ongoing disagreement among TS scholars regarding a standardized TQA model, as well as the disconnect between theoretical approaches and practical applications. Scholars and professionals in the field are still grappling with the challenge of developing a comprehensive and applicable model that reflects real-world demands.

Doubts and Directions: Emerging Research Strands

The partial failure of the proposed TQA models to meet real-world expectations has opened up venues for other research strands. In particular, two research strands are identifiable. In one strand, considering the criticisms leveled at TQA models, attempts have been made at developing new translation assessment grids to bridge the gap between the academic aspect of translation studies by incorporating concepts such as different facets of translation competence on the one hand, and the vocational expectations of the market place on the other. This strand seems to be particularly timely considering the failure of the existing models in meeting the needs of both educational and vocational settings, which ultimately should come to one another at a point of intersection. One such study is Orlando's (2011) work in designing two assessment grids for two types of assessment pursuing both pedagogical and vocational purposes. Asserting his belief as to the impossibility of finding any ideal system of evaluating translation, he holds that a system as such should be practical. Orlando's proposed grids are both product-oriented for summative evaluation and process-oriented for formative evaluation. What is promising with the grids is the fact that the grids have been in use for some years and feedback from both trainees and evaluators support their practicality and applicability.

Moreover, the failure seems to be an impetus for researchers to direct their attention away from the model and, instead, design research projects in order to test if it is only the model that plays a crucial role in determining the quality of translation. Supportive of this justification is Waddington's (2001) thought-provoking research. In an attempt to examine the validity of

current frequently-used methods of TQA in Spain, he brought four methods under scrutiny. The functioning of method A was based on the distinction between serious errors and minor errors while in method B translation and language mistakes were in focus. Method C was a holistic method and finally method D, as a hybrid method, was a combination of features of methods B and C. After running factor analysis, it was revealed that there were four factors underlying the 17 external criteria and Translation Competence turned out to be the most important factor. As for the methods of assessment, he came to the conclusion that all four currently used methods enjoyed certain degrees of criterion validity; he reported no significant difference between the validity of the methods.

Such studies as Waddington's reveal that it is not merely the model that plays a role in the disparities emerging from the outcome of TQAs. There have been some reflections on other possible sources paving the ground for the second emerging strand to form. It is argued that, as in any other decision-making process, personality traits affect the outcome.

Personality Traits and TQA

Personality traits refer to enduring patterns of thoughts, feelings, and behaviors that characterize an individual's unique personality (McCrae & Costa, 1999). They encompass a broad range of characteristics, including social attitudes, emotional stability, extraversion, conscientiousness, openness to experiences, and agreeableness.

There are various classifications for personality traits. The Five Factor Model, for instance, also known as the Big Five is one of the most widely accepted and extensively studied models of personality traits (Costa & McCrae, 1992). It comprises five main categories, inter alia, openness (to Experience), conscientiousness, extraversion, agreeableness, and finally neuroticism. "Openness to experience" reflects the individual's receptiveness to new ideas, imagination, and intellectual curiosity. "Conscientiousness" is defined as the degree of organization, responsibility, and dependability. Besides, "extraversion" is another trait describing the level of sociability, assertiveness, and positive emotional expression. The next trait, "agreeableness," indicates an individual's tendency to be cooperative, warm, and compassionate toward others. Finally, "neuroticism" represents the degree of emotional stability, anxiety proneness, and negative affect. The Myers-Briggs Type Indicator (MBTI), which is based on the work of Carl Jung (McCrae & Costa, 1999), and the HEXACO Model (Ashton & Lee, 2007) are among other personality inventories.

As stated above, one of the emerging research strands in TQA delineates the role of personality traits. While many have investigated the relationship between translators and trainees' personality traits and their translation quality (e.g., Gevaert, 2020; Hubscher-Davidson, 2013; Karimnia & Mahjubi, 2013; Lehka-Paul & Whyatt, 2016; Shaki & Khoshsaligheh, 2017), few studies have reflected on the raters' personality traits and their quality judgment. In one study, Conde (2012) directed his attention away from the model itself to the role the raters' personality traits can play by analyzing the behavior of demanding and lenient raters. The subjects were classified as demanding or lenient. The lenient raters of Cond's study worked more with the text, were more product-oriented, revealed a steadier performance, and were more confident and committed. Demanding raters, however, were less confident, were more feedback-oriented, and carried out fewer actions on the text. It was finally concluded that while demanding raters were better for professional purposes and advanced training, lenient raters seemed more appropriate for teaching at the initial stages.

Another study within the same research strand was conducted by Li and Yang (2023) who investigated the relationship between neuroticism, as one basic domain in the Big Five personality traits, and rater's performance in an interpreting context. They found that raters with a higher level of neuroticism tended to overestimate the problems and errors and, therefore, give lower scores. This finding corroborated Denovan, Dagnall, and Lofthouse's (2019) findings as they also found that among raters with high neuroticism scores, negative thinking and anxiety prevail, which, in turn, leads to their tendency to overestimate errors. Similarly, Buchanan (2017) and Uttl and Kibreab (2011) found that the neurotic personality of the rater can be a source of variance in quality judgment.

All in all, from what was reported above, it could be concluded that while most pronounced academic models of TQA, due to their drawbacks, have failed to be accepted as practical and applicable tools of assessment for both pedagogical and vocational purposes, a shift of direction is perceivable. This shift, as stated, seems to be two-fold: on the one hand, this shift seems to have had some effects on the way new tools of TQA have been designed as designing a complicated inapplicable model has been proved to be a tested and failed road; on the other hand, considering the assumption that the only key factor in TQA is the model is called into question, a vibrant research strand investigating the diverse range of personality traits that can affect raters' performance has been emerged.

This change of direction is welcome in research on TQA, especially considering that few pieces of academic studies could be found reflecting the new trend. The recent few studies are worth replicating in different contexts and testing to see their reliability as they have the

potential of marking a turning point in research in TQA. Considering the rise of TS as an academic discipline in the world, and accordingly establishment of translation degrees at both undergraduate and postgraduate levels at different Iranian universities, a pressing need is felt to conduct research studies whose findings help not only bridge the gap between vocational and educational settings but also improve the training programs and the way trainees are evaluated. Therefore, the present study with a view to highlighting the importance of this perceived shift, and in an attempt to shed light on the almost newly-emerged angels of TQA tries to address the following research questions:

(1) Is there any significant relationship between raters' personality traits of agreeableness and conscientiousness and leniency in translation assessment?

(2) Does the use of an agreed-upon TQA model help reduce leniency in translation assessment?

(3) Does the use of an agreed-upon TQA model affect the degree of inter-rater reliability?

(4) Is inter-rater reliability different among raters with different personality traits?

Method

Participants

Twenty-three senior B.A. students of TS were chosen through a non-random convenience sampling procedure. They were asked to translate an English excerpt with the specifications elaborated below. They were given the task as a take-home translation assignment and asked to submit their Persian translations within one week. The participants' age ranged from 20 to 28 with an average of 21. Among them, 16 were female and seven were male.

Seven translation instructors were asked to cooperate and assess students' translations. The instructors were all teaching translation at Iranian universities, with an average of eight years of teaching experience. Five were female and two were male.

Instruments

The following instruments were used to collect the data required:

Source Text

The source text given to the participants for translation was an authentic narrative extract of 270 words organized into four short paragraphs.

Big Five Inventory (BFI)

To gather information as to the raters' personality traits and specifically their extent of agreeableness and conscientiousness which are believed to be the predictors of leniency, the Big Five Inventory (BFI) was used. BFI is currently one of the commonplace instruments used in trait psychology when a short Big Five instrument is needed and when the focus is not on specific facets of personality. The reason why this instrument enjoys widespread use on the part of respondents can be attributed to its brevity, accessible vocabulary, and free availability. Its adequate level of internal consistency ($M\alpha$ = .83) makes it of interest in the eyes of researchers as well. Besides, the five scales of the BFI were shown to have convergent and discriminant validity with corresponding scales of other well-validated instruments such as Goldberg's descriptive adjectives and Costa and McCrae's 60-item NEO Five-Factor Inventory (John & Srivastava, 1999).

The BFI comprises five main categories, inter alia, openness (to Experience), conscientiousness, extraversion, agreeableness, and finally neuroticism among which conscientiousness and agreeableness are of interest in this study. Conscientiousness involves such constructs as perseverance, organized behavior, responsibility, thoroughness, and achievement-orientedness, and agreeableness is characterized by the tendency to be cooperative, considerate, and altruistic in social interactions (Costa & McCrae, 1992). The range of possible scores gained for conscientiousness and agreeableness are from 9 to 45 and 8 to 40, respectively with the lowest scores indicative of lesser degrees of conscientiousness and agreeableness and the highest scores suggesting a high extent of conscientiousness and agreeableness. All items are rated on a five-point scale, starting from strongly disagree to strongly agree.

Methods of Assessment

Each translated text was assessed by each rater twice. For the first time, no hints or guidelines were given to the raters as to how to assess; they were asked to assess the translations as they usually did, giving a final score between 0 to 20 to each translation as this range of scoring is the conventional one in the Iranian educational system.

However, for the second round of scoring an assessment tool was used. The tool was the one developed by Orlando (2011) for summative purposes. For this study, after discussing the applicability of the grid, minor modifications were made to it and, then, after arriving at a consensus, it was applied by the raters. The modifications made included taking out a part that assesses some ethical issues that were not relevant to this task. Adding to the weight of the first

parameters concerning the accuracy of comprehension and transfer was one further modification made and agreed upon among the raters (see Orlando (2011) for a more detailed description of how the tool is used.).

Table 1. Translation Assessment Grid

Translation Exercise	*/70*
• Comprehension of Source Text	0 2 4 6 8 10 12 14 16 18
(misinterpretations with more or less effect on accuracy)	
• Translation Accuracy / Transfer ST>TT	0 2 4 6 8 10 12 14 16
(mistranslations with more or less effect on accuracy)	
• Omissions / Insertions	0 2 4 6 8
(with more or less effect on accuracy)	
• Terminology / Word Choices	0 2 4 6 8
(affecting more or less the localized meaning)	
• Grammatical Choices / Syntactic Choices	0 2 4 6 8
(producing more or less distortion to the meaning)	
• Spelling Errors	0 2 4 6
• Punctuation Errors	0 1 2 3
• Formatting Errors	0 1 2 3
TRANSLATION EFFECT	*/30*
• Appropriateness for Target Audience	0 2 4 6 8 10
• Readability / Idiomatic Correctness	0 2 4 6 8 10
Adherence to Brief:	
• Function / Completeness	0 1 2 4 6
• Style / Presentation / Genre	0 1 2 4

Note. Adapted from Orlando's (2011) grid for summative translation assessment

Procedure

After collecting translations, the seven raters were asked to score translations separately. In doing so, no guidelines were given to them and they were free to assess the translations the way they usually did. The only prerequisite was to assess them all in one sitting. Two weeks after the last rater handed in the scored translations, the instructors were invited to discuss the usefulness of the pre-planned assessment grid (Orlando, 2011) and its applicability in educational settings. In the discussion session where discrepancies arose over the components of the grid, the problem was solved by agreement. When everyone agreed upon a final version,

the raters were asked to once again score each translation piece, which had been copied prior to the first scoring, this time using the agreed-upon grid. Like the previous situation, they were requested to do the whole job of assessment in one sitting.

As for the scales of conscientiousness and agreeableness, each rater was assigned one score, respectively based upon the data obtained from BFI. To address the research questions posed above, Pearson product-moment correlation coefficient was calculated to see if there was any relationship between raters' scores on the scales of agreeableness and conscientiousness on the one hand and leniency in assessment on the other, where leniency is defined as raters' tendency to assign ratings higher than justified by the real performance (Bernardin, Villanova, & Cook, 2000; Cascio & Aguinis, 2005; Tziner et al., 2005). The correlation coefficient was calculated for scorings, one in the absence of an agreed-upon model and one by the use of it. Furthermore, Cronbach's alpha was calculated to see if inter-rater reliability differed in the first scoring (without using the assessment tool) compared to the second round (with the use of the assessment tool). It also helped the researcher compare inter-rater reliability among more conscientious raters with less conscientious ones.

Results and Discussion

To investigate if there is a relationship between the raters' personality and leniency in assessment, each rater's score on the scales of "agreeableness" and "conscientious" was calculated. As mentioned earlier, the reason why these two scales were chosen to represent leniency is that research in psychology has recognized these two scales as predictors of leniency (Bernardin et al., 2000). Table 1 shows the raters' scores on each scale. As evident, under the headings of "agreeableness" and "conscientious," there are two columns: the first column (i.e., Total) shows the total score gained by each rater for the scale in question, and the second column (i.e., M) shows the average score of each rater with regard to that scale. The next two columns show the mean of the 23 scores given to the students' translations by each rater in the first and second rounds of scoring, respectively. Finally, the standard deviation and variance of the scores in the two rounds of scoring are reported.

Table 2. Descriptive Statistics

Rater	Agreeableness		Conscientiousness		M of Scores		SD of Scores	
	Total (40)	M (5)	Total (45)	M (5)	1st Round	2nd Round	1st Round	2nd Round
A	29	3.62	19	2.12	17.64	16.87	1.7	1.72
B	31	3.87	15	1.66	18.95	17.38	0.82	1.67
C	35	4.37	14	1.5	18.76	17.13	1.00	1.64
D	29	3.62	20	2.21	17.57	16.26	1.58	1.86
E	36	4.5	16	1.77	18.72	17.88	0.87	1.28
F	20	2.5	35	3.8	15.13	14.86	1.85	1.97
G	21	2.62	39	4.33	14.77	14.6	2.34	2.25

As Table 2 shows, two groups of raters emerged in the study: raters coded as A, B, C, and D seemed to be more agreeable and at the same time less conscientious, while raters F and G were more conscientious and less agreeable.

Personality and Leniency in Assessment

In what follows, the findings regarding the relationship between the personality traits of the raters and leniency in assessment in both the first and the second rounds of scoring are presented.

The First Round of Scoring

To investigate if there is a relationship between personality and leniency, the scores given by the raters were correlated with their scores in the two scales of "agreeableness" and "conscientiousness". As the analysis showed, the relationship between agreeableness and the scores given in the first round was positive and statistically significant ($r= 0.4$, $p<0.5$) which means the more agreeable the rater, the higher the scores s/he assigns to translations. In addition, conscientiousness revealed a high negative relationship with the first-round scores assigned ($r= -0.51$, $p<0.5$), indicating the tendency of more conscientious raters to assign lower scores to the performance of trainees –compared to less conscientious more agreeable raters– and to be far from leniency in their first round of scoring. This observation is consistent with the predicted hypothesis as to the relationship between these personality traits and leniency in translation assessment.

This finding is in line with the finding of Bernardin et al.'s (2000) study which suggests raters who are more agreeable and less conscientious make the most lenient and less accurate ratings. Bernardin et al.'s (2009) study has also come to the same conclusion. Though these studies are not on assessing translation in particular, they are psychological investigations about the way raters' traits might have an effect on their assessment job in general. Therefore, the studies as such are worth being investigated not only in different settings but also in various areas of assessment. The observation as to the tendency of more agreeable, less conscientious raters to assign higher scores to others' performance can be best justified considering the fact that those who scored high on agreeableness are less prone to find faults and be dependent, systematic, and self-effacing (Costa & McCrae, 1992; Kane et al., 1995). Research has also revealed that agreeable subjects are more in quest of social approval while trying to avoid conflict (Meier, Robinson, & Wilkowski, 2006). This is a rough combination of such features in the raters scored high on agreeableness which justifies their rating that seems to be higher than the actual performance. This observation can be even more truly justified considering the fact that in this study the five raters scored high on agreeableness and low on conscientiousness.

The Second Round of Scoring

The second research question concerned the application of one pre-determined agreed-upon assessment model with specifically set parameters and its possible effect on alleviating leniency. To this end, the raters' scores on the two personality traits were correlated with the second round of scores they assigned to the translations. As for the scale of agreeableness, although the Pearson Product Moment Correlation Coefficient still showed a relationship, the relationship was weak ($r= 0.31$, $p<0.5$). The value of the correlation between conscientiousness and the second round of scores, however, was not too different from what was found in the first round of scoring ($r= -0.53$, $p<0.5$), as in both a negative significant correlation was revealed.

As mentioned, the value of the observed correlation coefficient between agreeableness and leniency decreased in the second round of scoring where an assessment tool was used by the raters. This observation, however, was not evident among the conscientious raters. As a result, it was found that the use of an agreed-upon TQA grid moderates leniency in translation assessment. To justify this finding, we can resort to the notion of *accountability* as a factor that can moderate the predictive power of agreeableness and, therefore, leniency (Tziner et al., 2003). Accountability has been defined as "being answerable to audiences for performing up to certain prescribed standards, thereby fulfilling obligations, duties,

expectations, and other charges" (Schlenker et al., 1994, p. 634). To fit this notion to the present study and the observation made, it could be argued that in the second phase of scoring, the raters felt they should be more accountable for the ratings they made. As mentioned earlier, before raters assigned scores for the second time, they were introduced to an assessment tool and we all discussed the usability of the tool in the context. As a result, minor modifications were also made to the tool to fit our purpose. Finally, the raters were asked to assign scores to the copies of the same translations this time using the tool. The whole procedure made them aware of the fact that their assessment job is more serious than what they first had thought and their final assessments would be studied and analyzed; i.e., they felt they had to be much more accountable. Therefore, accountability, as Tziner et al. (2003) argue, has moderated the effect of agreeableness. Although, in the second round of scoring there was still a statistically significant relationship between the personality trait of agreeableness and the scores given by the raters, a sharp decrease was observed in the magnitude of the correlation coefficient. This decrease can be indicative of the higher accountability the raters felt in their second round of scoring. One more justification in support of what was observed in the second round is that the discussion session in which the tool was introduced and discussed made some of the raters aware of a number of parameters that should be paid attention to in the assessment. This could be a reason why the scores they assigned in the second phase were lower than those of the first phase.

As for more conscientious raters, it could be stated that the small difference observed between the value of the correlation coefficient between the first scoring and the second round confirms the idea that conscientious individuals tend to do their job the best way possible with the highest degree of accuracy. Accordingly, their sense of responsibility is an impetus for them to do their best even in situations where their performance is not being observed. Therefore, their sense of accountability in their assessment did not change their performance significantly in the second round compared with the first round. This is also in line with what Tziner et al. (2002) found in their psychological investigations; they came to the conclusion that raters who were high on conscientiousness were less influenced by the anticipation that they should be answerable to observers about their performance.

Inter-Rater Reliability

To address the third research question and see if the use of a TQA model could have an impact on the consistency of scores among the raters, regardless of their personality type, Cronbach's

alpha was calculated for both the first and second scorings. As reported in Table 3, the consistency of scores in the first round of scoring was not statistically significant (α= 0.58) revealing the fact that when raters scored the translations without any agreed-upon method, the scores one rater reported were not consistent with those reported by others. To see if the consistency among raters improved by the application of an agreed-upon unified model, Cronbach's alpha was calculated for the second round of scoring as well. This time, although the value of alpha increased (α= 0.69), it was not statistically significant either, showing a still low degree of inter-rater reliability. Therefore, in response to the third research question, it should be stated that the use of an agreed-upon model did not lead to an increase in the value of inter-rater reliability. The observed increase, though statistically insignificant, in the value of alpha can be best justified considering that in the second round of scoring, all raters were asked to score using the same tool and consequently the same parameters against which to assess the translations.

Table 3. Cronbach's α

Raters	Cronbach's alpha	
	1st Round	2nd Round
All	0.58	0.69
Category 1	0.37	0.48
Category 2	0.71	0.87

Note. Category 1 includes raters high on agreeableness and low on conscientiousness; Category 2 includes raters high on conscientiousness and low on agreeableness.

As mentioned earlier, the number of raters in the first category (those higher on agreeableness and lower on conscientiousness) was five whilst those in the second category (higher on the scale of conscientiousness and lower on agreeableness) were two. Therefore, it was probable that the recorded inconsistency among raters emanated mostly from the performance of the first category. To answer the fourth research question, inter-rater reliability was once again calculated, this time separately for each of the two categories of raters. While the value of Cronbach's alpha in the first round of scoring was 0.37 among the first category of raters, this value was proved to be much higher among the second category (α= 0.71). This confirmed the idea that the reported inconsistency had been mainly due to the performance of the first category of raters. Moreover, to see the effect of using an agreed-upon assessment tool on the performance of each category of raters, Cronbach's alpha of each category was

calculated for the second round as well. As the table shows, while this value was 0.48 for the first category, which is not an acceptable value for Cronbach's α, for the second category it was much higher (α= 0.87) showing a good amount of inter-rater reliability. Therefore, inter-rater reliability improved in the second round compared with the first scoring. However, the increase in the value among the first category of raters was not significant.

In short, inter-rater reliability among the first category of raters, who were more agreeable and less conscientious, was not acceptable and even the use of an agreed-upon model did not change this value significantly. In contrast, Cronbach's alpha showed a high value of inter-rater reliability for the second category of raters in both their first and second scorings. As a result, more conscientious less agreeable raters turned out to be more consistent in their ratings. This observation is supported by Waddington's (2001) study which revealed that the assiduous application of a model or any other agreed-upon or personal means of assessment is much more important than the model itself. Conscientious raters, as expected, are more accurate in and responsible for what is assigned to them. Therefore, even if no agreed-upon tool is given to them, they, consciously or unconsciously, follow certain guidelines while assessing. If, on the other hand, an assessment tool is given to them and they have agreed to assess the translations according to the parameters set in the tool, they certainly do their best to accurately apply the model. Therefore, in both situations, they try to be precise and this precision is the reason why they do more accurately and show more consistency in their performance. This was less evident among the first category of raters who were more agreeable and less conscientious.

Concluding Remarks

In this study, attempts were made to view translation quality assessment not only from the perspective of the models applied but also from the perspective of the personality traits of the raters. Therefore, the two personality traits of agreeableness and conscientiousness which are believed to be predictors of leniency were focused on and measured among the raters to see if there was a relationship between their judgments and their personality. The study revealed that more agreeable less conscientious raters, as predicted, were more lenient compared to the less agreeable more conscientious raters. It was also found that inter-rater reliability was high among more conscientious raters in both scorings, one in the absence of any agreed-upon model and according to the way they usually assessed translations, and another with the application of an agreed-upon assessment tool. This, however, turned out not to be the case

among the first category of raters who were high on agreeableness but low on conscientiousness.

This study, in alignment with the potentially new strand in translation quality assessment, brings the importance of factors other than the model of assessment to the fore, and suggests that while the presence of an agreed-upon model can help to reduce the subjectivity, the subjectivity in TQA is not only because of raters' having different parameters in mind for assessing the quality of a translation; the fluctuations are partly attributable to the personality traits of raters. The study bears implications for both educational settings and the workplace. Considering that each rater's dispositions can shape their performance in quality assessment, the study highlights the need for standardized assessment criteria incorporating this variable to mitigate potential biases arising from rater effects. It also points out the need to implement quality control measures. Moreover, awareness of raters' disposition can assist in identifying potential biases and facilitating training interventions to minimize the impact.

Further research should be conducted on more raters to see if the same results are confirmed. Besides, the validity of the findings should be tested in other settings since personality traits and their effects might be culture-dependent. It is also interesting to find out about the performance of those raters who do not easily fit into the two personality categories discussed in this study. As an example, raters who are low or high both on agreeableness and conscientiousness could be studied. Besides, further studies could be conducted on the way the variable of personality traits of raters can be incorporated into TQA tools.

References

Al-Qinai, J. (2000). Translation quality assessment: Strategies, parameters, and procedures. *Meta*, *45*(3), 497-519. https://doi.org/10.7202/001878ar

Ashton, M.C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150-166.

Bernardin, H. J., Tyler, C. L., & Villanova, P. (2009). Rating level and accuracy as a function of rater personality. *International Journal of Selection and Assessment*, *17*(3), 300-310. https://doi.org/10.1111/j.1468-2389.2009.00472.x

Bernardin, H. J., Villanova, P., & Cook, D. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology*, *85*(2), 232-236. https://doi.org/10.1037/0021-9010.85.2.232

Buchanan, T. (2017). Self-assessments of memory correlate with neuroticism and conscientiousness, not memory span performance. *Personality and Individual Differences, 105,* 19-23. https://doi.org/10.1016/j.paid.2016.09.031

Carrol, J. B. (1966). An experiment in evaluating the quality of translation. *Mechanical Translation, 9*(2), 55-66.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management*. Prentice Hall.

Conde, T. (2012). The good guys and the bad guys: The behavior of lenient and demanding translators. *Meta: Translators' Journal*, *57*(3), 763-86. https://doi.org/10.7202/1017090ar

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory and the NEO five-factor inventory*. Psychological Assessment Resources.

Denovan, A., Dagnall, N., & Lofthouse, G. (2019). Neuroticism and somatic complaints: Concomitant effects of rumination and worry. *Behavioral and Cognitive Psychotherapy, 47*(4), 431-445. https://doi.org/10.1017/S1352465818000619

Drugan, J. (2013). *Quality in professional translation: Assessment and improvement*. Bloomsbury Publishing.

Gevaert, M. (2020). *Personality in translation: An experimental study of the relationship between personality traits of student translators and translation quality* [Master's thesis, Ghent University].

Heaton, J. B. (1990). *Writing English language tests.* Longman.

House, J. (1997). *Translation quality assessment: A model revisited*. Gunter Narr Verlag.

House, J. (2001). Translation quality assessment: Linguistic description versus social evaluation. *Meta, 46*(2), 243-257. https://doi.org/10.7202/003141ar

House, J. (2009). *Translation.* Oxford University Press.

Hubscher-Davidson, S. (2013). The role of intuition in the translation process: A case study. *Translation and Interpreting Studies, 8*(2), 211-232. https://doi.org/10.1075/tis.8.2.05hub

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). The stability of rater leniency: Three studies. *Academy of Management Journal*, *38*(4), 1036–1051. https://doi.org/10.2307/256619

Karimnia, A., & Mahjubi, M. (2013). Individual differences and quality of translation: A personality-based perspective. *Psychology of Language and Communication*, *17*(1), 37-64. https://doi.org/10.2478/plc-2013-0003

Lehka-Paul, O., & Whyatt, B. (2016). Does personality matter in translation? Interdisciplinary research into the translation process and product. *Poznań Studies in Contemporary Linguistics*, *52*(2), 1-33. https://doi.org/10.1515/psicl-2016-0012

Li, H., & Yang, B. (2023). I misunderstand you because I worry about you: The relationship between neuroticism and ratings of linguistic interpreting. *Personality and Individual Differences*, *207*(8), 112-127. https://doi.org/10.1016/j.paid.2023.112170

McCrae, R. R., & Costa, P. T. (1999). A five-factor theory of personality. In L. A. Pervin and O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 139-153). Guilford Press.

Meier, B. P., Robinson, M. D., & Wilkowski, B. M. (2006). Turning the other cheek: Agreeableness and the regulation of aggression-related primes. *Psychological Science*, *17*(2), 136–142. https://doi.org/10.1111/j.1467-9280.2006.01676.x

Munday, J. (2016). *Introducing translation studies: Theories and applications* (4th ed.). Routledge.

Nida, E. (1964). *Towards a science of translating*. E. J. Brill.

Orlando, M. (2011). Evaluation of translations in the training of professional translators. *The Interpreter and Translator Trainer*, *5*(2), 293-308. https://dx.doi.org/10.1080/13556509.2011.10798822

Reiss, K. (1989). Text types, translation types, and translation assessment. Translated by A. Chesterman. In Chesterman, A. (ed.), *Readings in translation theory*. 105-115.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. *Psychological Review*, *101*(4), 632–652. https://doi.org/10.1037/0033-295x.101.4.632

Shaki, R., & Khoshsaligheh, M. (2017). Personality type and translation performance of Persian translator trainees. *Indonesian Journal of Applied Linguistics*, *7*(2), 360-370. http://dx.doi.org/10.17509/ijal.v7i2.8348

Toulmin, S. (1964). *The uses of argument*. Cambridge University Press.

Toulmin, S. (1984). *An introduction to reasoning*. Macmillan.

Tziner, A., Murphy, K. R., & Cleveland, J. (2002). Does conscientiousness moderate the relationship between attitudes and beliefs regarding performance appraisal and rating behavior? *International Journal of Selection and Assessment, 10*(3), 218–224. https://psycnet.apa.org/doi/10.1111/1468-2389.00211

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2003). Personality moderates the relationship between context factors and rating behavior. In: Shohov, S. P. (ed.), *Advances in Research Psychology*, Vol. 22. Nova.

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and rater factors affecting rating behavior. *Group and Organization Management, 30*(1), 89–98. https://psycnet.apa.org/doi/10.1177/1059601104267920

Uttl, B., & Kibreab, M. (2011). Self-report measures of prospective memory are reliable but not valid. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 65*(1), 57.

Waddington, C. (2001). Different methods of evaluating students' translations: The question of validity. *Meta, 46*(2), 311-325. https://doi.org/10.7202/004583ar

Williams, M. (2001). The application of argumentation theory to translation quality assessment. *Meta, 46*(2), 326-344.

Williams, M. (2005). *Translation quality assessment: An argumentation–centered approach*. University of Ottawa Press.

Williams, M. (2009). Translation quality assessment. *Mutatis Mutandis, 2*(1), 3-23. https://doi.org/10.1353/6617